Kernel construction

Manel Martínez-Ramón

ECE, UNM

October, 2018

- Mercer's theorem is one of the best known results of James Mercer (1883 – 1932).
- It has fundamental importance for kernel methods.
- It is the key idea behind the *kernel trick*, which allows to solve nonlinear optimization problems through the construction of kernelized counterparts of linear algorithms.
- The Mercer's Theorem can be stated as follows.

Theorem: (Mercer's Theorem, Aizerman et Al, 64)
Let $K(\mathbf{x}, \mathbf{x}')$ be a bivariate function fulfilling the Mercer condition, i.e.,

$$\int_{\mathbb{R}^{N_r} \times \mathbb{R}^{N_r}} f(\mathbf{x}) K(\mathbf{x}, \mathbf{x}') f(\mathbf{x}') d\mathbf{x} d\mathbf{x}' \geq 0$$

for any function such that

$$\int f^2(\mathbf{x}) d\mathbf{x} < \infty$$

.
Then, an RKHS $\mathcal{H}$ and a mapping function $\boldsymbol{\varphi}(\cdot)$, such that

$$K(\mathbf{x}, \mathbf{x}') = \langle \boldsymbol{\varphi}(\mathbf{x}), \boldsymbol{\varphi}(\mathbf{x}') \rangle$$

exist.

If we sample the integral, the inequality holds:

$$\int_{\mathbb{R}^{N_r}\times\mathbb{R}^{N_r}} f(\mathbf{x})K(\mathbf{x},\mathbf{x}')f(\mathbf{x}')d\mathbf{x}d\mathbf{x}' \geq 0 \Leftrightarrow \sum_{i,j=1}^{N} f(\mathbf{x}_i)K(\mathbf{x}_i,\mathbf{x}_j)f(\mathbf{x}_j) \geq 0$$

With the change of notation $f(\mathbf{x}_i) = \alpha_i$ we can say that $K(\mathbf{x}_i,\mathbf{x}_j)$ is a dot product in a given $\mathcal{H}$ if and only if

$$\sum_{i,j=1}^{N} \alpha_i K(\mathbf{x}_i,\mathbf{x}_j)\alpha_j = \boldsymbol{\alpha}^\top \mathbf{K}\boldsymbol{\alpha} \geq 0$$

Some more properties of kernels

The following are called the closure properties of kernels, and allows us to produce new kernels from combinations of simple ones.

- Property 1 (*direct sum of Hilbert spaces*): The linear combination

$$k(\mathbf{x}, \mathbf{z}) = ak_1(\mathbf{x}, \mathbf{z}) + bk_2(\mathbf{x}, \mathbf{z})$$

  where $a, b \geq 0$ is a kernel.

- Proof: Let $\mathbf{x}_1, \cdots, \mathbf{x}_N$ be a set of points, and $\mathbf{K}_1$ and $\mathbf{K}_2$ the corresponding kernel matrices constructed with $k_1(\cdot, \cdot)$ and $k_2(\cdot, \cdot)$. Since these matrices are definite positive, so is $\mathbf{K}$ constructed with $k(\cdot, \cdot)$. Indeed, for any vector $\boldsymbol{\alpha} \in \mathbb{R}^N$

$$\boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha} = a\boldsymbol{\alpha}^\top \mathbf{K}_1 \boldsymbol{\alpha} + b\boldsymbol{\alpha}^\top \mathbf{K}_2 \boldsymbol{\alpha} \geq 0$$

The previous property can be also proved as follows:

- Proof: Let $\boldsymbol{\varphi}_1(\mathbf{x})$ and $\boldsymbol{\varphi}_2(\mathbf{x})$ be transformations to the RKHS spaces $\mathcal{H}_1$ and $\mathcal{H}_2$ respectively endowed with dot products $k_1(\cdot,\cdot)$ and $k_2(\cdot,\cdot)$.
- A vector in a composite or embedded Hilbert Space $\mathcal{H}$ can be constructed as

$$\boldsymbol{\varphi}(\mathbf{x}) = \begin{pmatrix} \sqrt{a}\boldsymbol{\varphi}_1(\mathbf{x}) \\ \sqrt{b}\boldsymbol{\varphi}_2(\mathbf{x}) \end{pmatrix}$$

Some more properties of kernels

- The corresponding kernel $k(\mathbf{x}, \mathbf{z})$ in space $\mathcal{H}$ is

$$
\begin{aligned}
k(\mathbf{x}, \mathbf{z}) &= \begin{pmatrix} \sqrt{a}\boldsymbol{\varphi}_1(\mathbf{x}) \\ \sqrt{b}\boldsymbol{\varphi}_2(\mathbf{x}) \end{pmatrix}^\top \begin{pmatrix} \sqrt{a}\boldsymbol{\varphi}_1(\mathbf{z}) \\ \sqrt{b}\boldsymbol{\varphi}_2(\mathbf{z}) \end{pmatrix} \\
&= a\boldsymbol{\varphi}_1^\top(\mathbf{x})\boldsymbol{\varphi}_1(\mathbf{z}) + b\boldsymbol{\varphi}_2^\top(\mathbf{x})\boldsymbol{\varphi}_2(\mathbf{z}) \\
&= ak_1(\mathbf{x}, \mathbf{z}) + bk_2(\mathbf{x}, \mathbf{z})
\end{aligned}
$$

- Hence, the linear combination of two kernels correspond to a kernel in a space $\mathcal{H}$ that embeds the corresponding RKHSs of both kernels.
- This is often called *direct sum of Hilbert spaces.*

- Property 2 (*Tensor product of kernels*): The product of kernels

$$k(\mathbf{x}, \mathbf{z}) = k_1(\mathbf{x}, \mathbf{z}) \cdot k_2(\mathbf{x}, \mathbf{z})$$

  is a kernel.

- Proof: Let $\mathbf{K} = \mathbf{K}_1 \otimes \mathbf{K}_2$ be the tensor product between kernel matrices, where each element $k_1(\mathbf{x}_i, \mathbf{x}_j)$ of matrix $\mathbf{K}_1$ is replaced by the product $k_1(\mathbf{x}_i, \mathbf{x}_j)\mathbf{K}_2$.
- The eigenvalues of the tensor product are all the products of eigenvalues of both matrices. Then, for any $\boldsymbol{\alpha} \in \mathbb{R}^{N \cdot N}$

$$\boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha} \geq 0$$

Some more properties of kernels

- In particular, the Schur product matrix $\mathbf{H}$ with entries $H_{i,j} = k_1(\mathbf{x}_i, \mathbf{x}_j)$ is a submatrix of $\mathbf{K}$ defined by a set of columns and the same set of rows. Assume that a vector $\boldsymbol{\alpha}$ exists with nonnull elements in these positions, and zero in the rest. Then

$$\boldsymbol{\alpha}^\top \mathbf{K}\boldsymbol{\alpha} = \boldsymbol{\alpha}'^\top \mathbf{H}\boldsymbol{\alpha}' \geq 0$$

where $\boldsymbol{\alpha}' \in \mathbb{R}^N$ is the vector constructed with the nonnull components of $\boldsymbol{\alpha}$.

Some more properties of kernels

- Property 3: $k(\mathbf{x}, \mathbf{z}) = f(\mathbf{x}) \cdot f(\mathbf{z})$ is a kernel. Straightforwardly, function $f(\mathbf{x})$ is a one dimensional map to $\mathbb{R}$.
- Property 4. $k(\boldsymbol{\varphi}(\mathbf{x}), \boldsymbol{\varphi}(\mathbf{z}))$ is a kernel.
- Property 5. If $\mathbf{B}$ is positive semi definite, then $\mathbf{x}^\top \mathbf{B} \mathbf{z}$ is a kernel.
- **Exercise**. Determine in what cases the product $\mathbf{x}^\top \mathbf{B} \mathbf{z}$ with $\mathbf{X} \in \mathbb{R}^{D_1}$, $\mathbf{X} \in \mathbb{R}^{D_2}$ and $\mathbf{B}$ has dimensions $D_1 \times D_2$. Hint: use eigendecomposition

Kernel construction

Manel Martínez-Ramón

ECE, UNM

October, 2018

Yet some more properties of kernels

Let $k_1(\mathbf{x}, \mathbf{z})$ be a kernel. The following are also kernels:

1. $k(\mathbf{x}, \mathbf{z}) = p\left(k_1(\mathbf{x}, \mathbf{z})\right)$ where $p(v)$ is a polynomial function of $v \in \mathbb{R}$ with positive coefficients.
2. $k(\mathbf{x}, \mathbf{z}) = \exp(k_1(\mathbf{x}, \mathbf{z}))$
3. $k(\mathbf{x}, \mathbf{z}) = \exp\left(\frac{-||\mathbf{x}-\mathbf{z}||^2}{2\sigma^2}\right)$
   where
   $$||\mathbf{x} - \mathbf{z}||^2 = \langle \boldsymbol{\varphi}(\mathbf{x}) - \boldsymbol{\varphi}(\mathbf{z}), \boldsymbol{\varphi}(\mathbf{x}) - \boldsymbol{\varphi}(\mathbf{z}) \rangle$$

Proofs:

1. By property 2, $k_p(\mathbf{x}, \mathbf{z}) = k(\mathbf{x}, \mathbf{z})^p$, where $p \in \mathbb{N}$, is a kernel. Then, by property 1 we can say that

$$k(\mathbf{x}, \mathbf{z}) = \sum_{p=1}^{P} a_p k(\mathbf{x}, \mathbf{z})^p + a_0$$

   where $a_p \geq 0$, is a kernel.

Proofs:

3. The Taylor series expansion of the exponential is

$$\exp(v) = \sum_{k=0}^{\infty} \frac{1}{k!} v^k$$

It is a polynomial with positive coefficients, hence a kernel.

Proofs:

4. We can expand the norm of a distance vector as

$$\begin{aligned}||\mathbf{x}-\mathbf{z}||^2 &= \langle \boldsymbol{\varphi}(\mathbf{x})-\boldsymbol{\varphi}(\mathbf{z}), \boldsymbol{\varphi}(\mathbf{x})-\boldsymbol{\varphi}(\mathbf{z})\rangle \\ &= k(\mathbf{x},\mathbf{x})+k(\mathbf{z},\mathbf{z})-2k(\mathbf{x},\mathbf{z})\end{aligned}$$

The squared exponential of this norm is

$$\begin{aligned}k(\mathbf{x},\mathbf{z}) &= \exp\left(\frac{-||\mathbf{x}-\mathbf{z}||^2}{2\sigma^2}\right) \\ &= \exp\left(-\frac{k(\mathbf{x},\mathbf{x})}{2\sigma^2}-\frac{k(\mathbf{z},\mathbf{z})}{2\sigma_2}+\frac{k(\mathbf{x},\mathbf{z})}{\sigma^2}\right) \\ &= \frac{\exp\left(\frac{k(\mathbf{x},\mathbf{z})}{\sigma^2}\right)}{\exp\left(\frac{k(\mathbf{x},\mathbf{x})}{2\sigma^2}\right)\exp\left(\frac{k(\mathbf{z},\mathbf{z})}{2\sigma_2}\right)}\end{aligned}$$

$$
\begin{aligned}
&= \frac{\exp\left(\frac{k(\mathbf{x},\mathbf{z})}{\sigma^2}\right)}{\exp\left(\frac{k(\mathbf{x},\mathbf{x})}{2\sigma^2}\right)\exp\left(\frac{k(\mathbf{z},\mathbf{z})}{2\sigma_2}\right)} \\
&= \frac{\exp\left(\frac{k(\mathbf{x},\mathbf{z})}{\sigma^2}\right)}{\sqrt{\exp\left(\frac{k(\mathbf{x},\mathbf{x})}{\sigma^2}\right)\exp\left(\frac{k(\mathbf{z},\mathbf{z})}{\sigma_2}\right)}} \\
&= \frac{\kappa(\mathbf{x},\mathbf{z})}{\sqrt{\kappa(\mathbf{x},\mathbf{x})\kappa(\mathbf{z},\mathbf{z})}}
\end{aligned}
$$

Since by previous property $\kappa(\mathbf{x},\mathbf{z}) = \exp\left(\frac{k(\mathbf{x},\mathbf{z})}{\sigma^2}\right)$ is a kernel, this expression is a (normalized) kernel, since it is also positive semidefinite.

Kernels in closed form

$$\begin{aligned}
\text{Linear:} \quad & k(\mathbf{x},\mathbf{y}) = \mathbf{x}^\top \mathbf{y} + c \\
\text{Polynomial:} \quad & k(\mathbf{x},\mathbf{y}) = (\alpha \mathbf{x}^\top \mathbf{y} + c)^d \\
\text{Gaussian:} \quad & k(\mathbf{x},\mathbf{y}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}\right) \\
\text{Exponential:} \quad & k(\mathbf{x},\mathbf{y}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|}{2\sigma^2}\right) \\
\text{Laplacian:} \quad & k(\mathbf{x},\mathbf{y}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|}{\sigma}\right)
\end{aligned}$$

Kernels in closed form

$$\text{Hyperbolic Tangent (Sigmoid):} \quad k(\mathbf{x},\mathbf{y}) = \tanh(\alpha \mathbf{x}^\top \mathbf{y} + c)$$

$$\text{Rational Quadratic:} \quad k(\mathbf{x},\mathbf{y}) = 1 - \frac{(\mathbf{x}-\mathbf{y})^2}{(\mathbf{x}-\mathbf{y})^2 + c}$$

$$\text{Multiquadric:} \quad k(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})^2 + c^2}$$

$$\text{Inverse Multiquadric:} \quad k(\mathbf{x},\mathbf{y}) = \frac{1}{\sqrt{(\mathbf{x}-\mathbf{y})^2 + \theta^2}}$$

Kernels in closed form

$$\text{Power:} \quad k(\mathbf{x}, \mathbf{y}) = -(x - y)^d$$

$$\text{Log:} \quad k(\mathbf{x}, \mathbf{y}) = -\log((\mathbf{x} - \mathbf{y})^d + 1)$$

$$\text{Cauchy:} \quad k(\mathbf{x}, \mathbf{y}) = \frac{1}{1 + \frac{(\mathbf{x}-\mathbf{y})^2}{\sigma^2}}$$

$$\text{Chi-Square:} \quad k(\mathbf{x}, \mathbf{y}) = 1 - \sum_{k=1}^{d} \frac{(\mathbf{x}^{(k)} - \mathbf{y}^{(k)})^2}{\frac{1}{2}(\mathbf{x}^{(k)} + \mathbf{y}^{(k)})}$$

$$\text{Histogram (or min) Intersection:} \quad k(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^{d} \min(\mathbf{x}^{(k)}, \mathbf{y}^{(k)})$$

Kernels in closed form

$$\text{Generalized Hist. Intersection:}\ \ k(\mathbf{x},\mathbf{y}) = \sum_{k=1}^{m} \min(|\mathbf{x}^{(k)}|^{\alpha}, |\mathbf{y}^{(k)}|^{\beta})$$

$$\text{Generalized } T\text{-Student:}\ \ k(\mathbf{x},\mathbf{y}) = \frac{1}{1+(\mathbf{x}-\mathbf{y})^{d}}$$

The book "Kernel Methods for Pattern Analysis" by J. Shawe-Taylor and N. Cristianini is a comprehensive document in this topic. Please take an online look at it from the UNM library. It is part of the reference documents of this class.

Outcomes of this lesson

We have seen the Mercer's theorem, a property that is fundamental to kernel definition and for its use in Machine Learning. This allows to justify the construction of kernels. We have reviewed the following kernels:

- Sum of kernels (slide 5)
- Products of kernels (slide 8)
- Kernels as product of functions (slide 10)
- Kernels embedded in kernels (slide 11)
- Polynomials odf kernels (slide 12)
- Some closed form kernels (slide 16)